

Preserving Individual Privacy in Serial Data Publishing

Raymond Chi-Wing Wong¹, Ada Wai-Chee Fu², Jia Liu², Ke Wang³, Yabo Xu³

¹ Hong Kong University of Science and Technology ²Chinese University of Hong Kong

raywong@cse.ust.hk

{adafu,jliu}@cse.cuhk.edu.hk

³ Simon Fraser University

{wangk,yxu}@cs.sfu.ca

ABSTRACT

While previous works on privacy-preserving serial data publishing consider the scenario where sensitive values may persist over multiple data releases, we find that no previous work has sufficient protection provided for sensitive values that can change over time, which should be the more common case. In this work, we propose to study the privacy guarantee for such transient sensitive values, which we call the *global guarantee*. We formally define the problem for achieving this guarantee and derive some theoretical properties for this problem. We show that the anonymized group sizes used in the data anonymization is a key factor in protecting individual privacy in serial publication. We propose two strategies for anonymization targeting at minimizing the average group size and the maximum group size. Finally, we conduct experiments on a medical dataset to show that our method is highly efficient and also produces published data of very high utility.

1. INTRODUCTION

Recently, there has been much study on the issues in privacy-preserving data publishing [2, 13, 12, 4, 16, 27, 7, 14, 33, 9, 22, 15]. Most previous works deal with privacy protection when only one instance of the data is published. However, in many applications, data is published at regular time intervals. For example, the medical data from a hospital may be published twice a year. Some recent papers [19, 30, 8, 6, 23, 5] study the privacy protection issues for *multiple* data publications of multiple instances of the data. We refer to such data publishing *serial data publishing*.

Following the settings of previous works, we assume that there is a sensitive attribute which contains sensitive values that should not be linked to the individuals in the database. A common example of such a sensitive attribute is diseases. While some diseases such as flu or stomach virus may not be very sensitive, some diseases such as chlamydia (a sex disease) can be considered highly sensitive. In serial publishing of

such a set of data, the disease values attached to a certain individual can change over time.

A typical guarantee we want to achieve is that the probability that an adversary can derive for the linkage of a person to a sensitive value is no more than $1/\ell$. This is well-known to be a simple form of ℓ -diversity [16]. This guarantee sounds innocent enough for a single release data publication. However, when it comes to serial data publishing, the objective becomes quite illusive and requires a much closer look. In serial publishing, the individuals that are recorded in the data may change, and the sensitive values related to individuals may also change. We assume that the sensitive values can change freely.

Let us consider a sensitive disease chlamydia, which is a sex disease that is easily curable. Suppose that there exist 3 records of an individual o in 3 different medical data releases. It is obvious that typically o would not want anyone to deduce with high confidence from these released data that s/he has ever contracted chlamydia in the past. Here, the past practically corresponds to *one or more* of the three data releases. Therefore, if from these data releases, an adversary can deduce with high confidence that o has contracted chlamydia in one or more of the three releases, privacy would have been breached. To protect privacy, we would like the probability of any individual being linked to a sensitive value in one or more data releases to be bounded from the above by $1/\ell$. Let us call this privacy guarantee the *global guarantee* and the value $1/\ell$ the *privacy threshold*.

Though the global guarantee requirement seems to be quite obvious, to the best of our knowledge, no existing work has considered such a guarantee. Instead, the closest guarantee of previous works is the following: for *each* of the data releases, o can be linked to chlamydia with a probability of no more than $1/\ell$. Let us call this guarantee the *localized guarantee*. Would this guarantee be equivalent to the above global guarantee? In order to answer this question, let us look at an example.

Consider two raw medical tables (or micro data) T_1 and T_2 as shown in Figure 1 at time points 1 and 2, respectively. Suppose that they contain records for the individuals o_1, o_2, o_3, o_4, o_5 . There are two kinds of attributes, namely *quasi-identifier* (*QID*) attributes and *sensitive* attributes. Quasi-identifier attributes are attributes that can be used to identify an individual with the help of an external source such as a voter registration list [21, 12, 13, 29]. In this example, sex and zipcode are the quasi-identifier attributes, while disease is the sensitive attribute. Attribute id is used for illustration purpose and does not appear in the published table. We as-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

sume that each individual owns at most one tuple in each table at each time point. Furthermore, we assume no additional background knowledge about the linkage of individuals to diseases, and the sensitive values linked to individuals can be freely updated from one release to the next release.

Id	Sex	Zip-code	Disease
o_1	M	65001	flu
o_2	M	65002	chlamydia
o_3	F	65014	flu
o_4	F	65015	fever

(a) T_1

Id	Sex	Zip-code	Disease
o_1	M	65001	chlamydia
o_2	M	65002	flu
o_3	F	65014	fever
o_5	F	65010	flu

(b) T_2

Figure 1: A motivating example

Sex	Zipcode	Disease
M	6500*	flu
M	6500*	chlamydia
F	6501*	flu
F	6501*	fever

(a) T_1^*

Sex	Zipcode	Disease
M	6500*	chlamydia
M	6500*	flu
F	6501*	fever
F	6501*	flu

(b) T_2^*

Figure 2: Anonymization for T_1 and T_2

Assume that the privacy threshold is $1/\ell = 1/2$. In a typical data anonymization [21, 12, 13, 29], in order to protect individual privacy, the QID attributes of the raw table are *generalized* or *bucketized* in order to form some *anonymized groups* (\mathcal{AG}) to hide the linkage between an individual and a sensitive value. For example, table T_1^* in Figure 2(a) is a *generalized* table of T_1 in Figure 1. We generalize the zip code of the first two tuples to 6500* so that they have the same QID values in T_1^* . We say that these two tuples form an *anonymized group*. It is easy to see that in each published table T_1^* or T_2^* , the probability of linking any individual to chlamydia or flu is at most $1/2$, which satisfies the localized guarantee. The question is whether this satisfies the global privacy guarantee with a threshold of $1/2$.

For the sake of illustration, let us focus on the anonymized groups G_1 and G_2 containing the first two tuples in tables T_1^* and T_2^* in Figure 2, respectively. The probability in serial publishing can be derived by the possible world analysis. There are four possible *worlds* for G_1 and G_2 in these two published tables, as shown in Figure 3. Here each *possible world* is one possible way to assign the diseases to the individuals in such a way that is consistent with the published tables. Therefore, each possible world is a possible assignment of the sensitive values to the individuals at all the publication time points for groups G_1 and G_2 . Note that an individual can be assigned to different values at different data releases, and the assignment in one data release is independent of the assignment in another release.

Consider individual o_2 . Among the four possible worlds, three possible worlds link o_2 to “chlamydia”, namely w_1, w_2 and w_3 . In w_1 and w_2 , the linkage occurs at T_1 , and in w_3 , the linkage occurs at T_2 . Thus, the probability that o_2 is linked to “chlamydia” in at least one of the tables is equal to $3/4$, which is greater than $1/2$, the intended privacy threshold. From this example, we can see that localized guarantee does not imply global guarantee.

In this paper, we show that in order to ensure the global guarantee, the sizes of the anonymized groups need to be bigger than that needed for localized guarantee. In the above

Sex	Zipcode	Disease
M/F	650**	flu
M/F	650**	chlamydia
M/F	650**	flu
M/F	650**	fever

(a) T_1^*

Sex	Zipcode	Disease
M/F	650**	chlamydia
M/F	650**	flu
M/F	650**	fever
M/F	650**	flu

(b) T_2^*

Figure 4: Anonymization for global guarantee

example, we can use size 4 anonymized groups as shown in Figure 4. There will be $4! \times 4!$ possible worlds. It is easy to see that $3/4$ of the possible worlds do not assign chlamydia to o_2 in the first release, $3/4$ of them do not assign chlamydia to o_2 in the second release, and $3/4 \times 3/4 = 9/16$ of the possible worlds do not assign chlamydia to o_2 in both releases. The remaining possible worlds assign chlamydia to o_2 in at least one of the two releases. Hence, the privacy breach probability $= 1 - 9/16 = 7/16 < 1/2$.

The contributions of this paper include the following: We point out the problem of privacy breach that arises with localized guarantee and propose to study the problem of global guarantee in privacy preserving serial data publishing. We formally analyze the privacy breach with transient sensitive values. Useful properties related to the anonymization under the global guarantee are derived. These properties are related to the anonymized group sizes. Typically group sizes greater than that required for the localized guarantee will be needed to attain the global guarantee. These properties are then leveraged in the proposal of new anonymization strategies that can minimize the information loss. We have also conducted extensive experiments with a real medical dataset to verify our techniques. The results show that our methodology are very promising in real world applications.

The rest of this paper is organized as follows. Section 2 surveys the previous related works. Section 3 contains our problem definition. Section 4 describes a general formula for the breach probability. Section 5 discusses some key properties for this problem. Section 6 describes our methodology for privacy protection. Section 7 suggests a possible implementation. Section 8 is an empirical study. Section 9 concludes our work and points out some possible future directions.

2. RELATED WORK

Here, we summarize the previous works on the problem of privacy preserving serial data publishing. k -anonymity has been considered in [8] and [19] for serial publication allowing only insertions, but they do not consider the linkage probabilities to sensitive values. The work in [23] considers sequential releases for different attribute subsets for the same dataset, which is different from our definition of serial publishing.

There are some more related works that attempt to avoid the linkage of individuals to sensitive values. Delay publishing is proposed in [6] to avoid problems of insertions, but deletion and updates are not considered. While [30] considers both insertions and deletions, both [6] and [30] make the assumption that when an individual appears in consecutive data releases, then the sensitive value for that individual is not changed. As pointed out in [5], this assumption is not realistic. Also the protection in [30] is record-based and not individual-based. This is quite problematic, as in our running examples, there are two *records* for one *individual* o_2 , namely, t_1 in table T_1 and t_2 in table T_2 (note that T_1 and T_2 need not be con-

Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia
T_1		
Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia
T_1		
Sex	Zipcode	Disease
M	65001	flu
M	65002	flu
T_2		
Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu
T_2		

(a) Possible world 1 w_1

Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia
T_1		
Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu
T_2		

(b) Possible world 2 w_2

Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu
T_1		
Sex	Zipcode	Disease
M	65001	flu
M	65002	chlamydia
T_2		

(c) Possible world 3 w_3

Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu
T_1		
Sex	Zipcode	Disease
M	65001	chlamydia
M	65002	flu
T_2		

(d) Possible world 4 w_4

Figure 3: Possible worlds for G_1 and G_2

secutive releases, so that the sensitive value linked to o_2 can change even if we adopt the above unrealistic assumption in [6, 30]). If we consider just tuple t_1 , then there are only 2 possible worlds where t_1 is linked to chlamydia in Figure 3, namely w_1 and w_2 . If we just consider tuple t_2 , there are also only 2 possible worlds linking it to chlamydia, namely w_1 and w_3 . Hence, T_1^* and T_2^* satisfy the record-based requirement of [30] if the risk threshold is 0.5. In fact, these are possible tables generated by the mechanism proposed in [30]. However, we have shown that this anonymization does not provide the expected protection for the individuals.

The ℓ -scarcity model is introduced in [5] to handle the situations when some data may be permanent so that once an individual is linked to such a value, the linkage will remain in subsequent releases whenever the individual appears (not limited to consecutive releases only). However, for transient sensitive values, [30] and [5] adopt the following principle.

PRINCIPLE 1 (LOCALIZED GUARANTEE). *For each release of the data publication, the probability that an individual is linked to a sensitive value is bounded by a threshold.*

However, we have seen in the example in the previous section that this cannot satisfy the expected privacy requirement. Hence, we consider the following principle.

PRINCIPLE 2 (GLOBAL GUARANTEE). *Over all the published releases, the probability that an individual has ever been linked to a sensitive value is bounded by a threshold.*

Although the privacy guarantee is the most important data publication criterion, the published data must also provide a reasonable level of utility so that it can be useful for applications such as data mining or data analysis. Utility is a tradeoff for the privacy guarantee since anonymization of data introduces information loss. There are different definitions of utility in the existing literature. Here, we briefly describe some common definitions.

The anonymized group sizes have been considered in utility metrics. The average group size is considered in [16]. In [3], the discernability model assigns a penalty to each tuple t as determined by the square of the size of the anonymized group for t . In [12], the normalized average anonymized group size metric is proposed, which is given by the total number of tuples in the table divided by the product of the total number of anonymized groups and a value k (for k -anonymity). Here, the best case occurs when each group has size k .

Other works [29, 31, 26] consider categorical data that comes with a taxonomy so that the information loss is measured with respect to the structure in the taxonomy when data are generalized from the leaf nodes to higher levels in the taxonomy. Both [11] and [28] measure utility by comparing the data distributions before and after anonymization.

Recently, [20] and [33] consider the accuracy in answering aggregate queries to be a measure of utility.

[10, 24, 1] assume that the data is utilized for classification and hence define the utility accordingly. The anonymization mechanisms in [17, 2, 32] are by means of suppressing data entries in the table, and hence information loss is measured by the number of suppressed entries.

3. PROBLEM DEFINITION

Suppose tables T_1, T_2, \dots, T_k are generated at time points, $1, 2, \dots, k$, respectively. Each table T_i has two kinds of attributes, *quasi-identifier attributes* and *sensitive attributes*. For the sake of illustration, we consider one single sensitive attribute S containing $|S|$ values, namely $s_1, s_2, \dots, s_{|S|}$. Assume that the sensitive values for individuals can freely change from one release to another release so that the linkage of an individual o to a sensitive value s in one data release has no effect on the linkage of o to any other sensitive value in any other data release. Assume at each time point j , a data publisher generates an anonymized version T_j^* of T_j for data publishing so that each record in T_j will belong to one anonymized group G in T_j^* . Given an anonymized group G , we define $G.S$ to be a multi-set containing all sensitive values in G , and $G.I$ to be the set of individuals that appear in G .

DEFINITION 1 (POSSIBLE WORLD). *A series of tables $TS = \{T_1^p, T_2^p, \dots, T_k^p\}$ is a possible world for published tables $\{T_1^*, T_2^*, \dots, T_k^*\}$ if the following requirement is satisfied. For each $i \in [1, k]$,*

1. *there is a one-to-one corresponding between individuals in T_i^p and individuals in T_i^**
2. *for each anonymized group G in T_i^* , the multi-set of the sensitive values of the corresponding individuals in T_i^p is equal to $G.S$.*

Let $p(o, s, k)$ be the probability that an individual o is linked to s in at least one published table among published tables $T_1^*, T_2^*, \dots, T_k^*$.

Let $t.S$ stand for the sensitive value of tuple t . We say that o is linked to s in a table T_i^p if for the tuple t of o in T_i^p , $t.S = s$. Following previous works, we define the probability based on the *possible worlds* as follows.

DEFINITION 2 (BREACH PROBABILITY). *The breach probability is given by*

$$p(o, s, k) = \frac{W_{link}(o, s, k)}{W_{total, k}} \quad (1)$$

where $W_{link}(o, s, k)$ is the total number of possible worlds where o is linked to s in at least one published table among

$T_1^p, T_2^p, \dots, T_k^p$ and $W_{total,k}$ is the total number of possible worlds for published tables $T_1^*, T_2^*, \dots, T_k^*$.

We will describe how we derive a general formula to calculate $p(o, s, k)$ in Section 4.

While privacy breach is the most important concern, the utility of the published data also need to be preserved. There are different definitions of utility in the existing literature. Some commonly adopted utility measurements are described in Section 2.

In this paper, we are studying the following problem.

PROBLEM 1. *Given a privacy parameter ℓ (a positive integer), a utility measurement, $k - 1$ published tables, namely $T_1^*, T_2^*, \dots, T_{k-1}^*$ and one raw table T_k , we want to generate a published table T_k^* from T_k such that the utility is maximized, and for each individual o and each sensitive value s ,*

$$p(o, s, k) \leq 1/\ell$$

Note that the above problem definition follows Principle 2 for global guarantee as discussed in Section 2.

3.1 Global versus Localized Guarantee

Here, we show that protecting individual privacy with Principle 2 (global guarantee) implies protecting individual privacy with Principle 1 (localized guarantee). Under Principle 1, let $q(o, s, j, k)$ be the probability that an individual o is linked to a sensitive value s in the j -th table. Following the definition of probability adopted in most previous works [30, 5], we have

$$q(o, s, j, k) = \frac{L_{link}(o, s, j, k)}{W_{total,k}}$$

where $L_{link}(o, s, j, k)$ is the total number of possible worlds in which o is linked to s in the j -th table and $W_{total,k}$ is the total number of possible worlds for the k published tables.

In our running example, $k=2$ and from Figure 3, there are four possible worlds, $W_{total,k} = 4$. Consider published table T_1^* . There are two possible worlds where o_2 is linked to chlamydia (s), namely w_1 and w_2 . Thus, $L_{link}(o_2, s, 1, k) = 2$ and $q(o_2, s, 1, k) = \frac{2}{4} = \frac{1}{2}$. Similarly, when $j = 2$, $q(o_2, s, 2, k) = \frac{1}{2}$.

In general, it is obvious that $W_{link}(o, s, k) \geq L_{link}(o, s, j, k)$ for any $j \in [1, k]$. We derive that

$$p(o, s, k) \geq q(o, s, j, k)$$

Hence we have the following lemma.

LEMMA 1. *If $p(o, s, k) \leq 1/\ell$ (under Principle 2), then for any $j \in [1, k]$, $q(o, s, j, k) \leq 1/\ell$ (under Principle 1).*

COROLLARY 1. *Principle 2 (global guarantee) is a strictly stronger requirement than Principle 1 (localized guarantee).*

4. BREACH PROBABILITY ANALYSIS

In this section, we consider how the breach probability $p(o, s, k)$ can be derived. For privacy breach, we focus on the possible assignment of sensitive values to one individual at a time. Therefore, we introduce the following possible world definition to deal with assignments to a particular individual.

DEFINITION 3 (\mathcal{AG}_i). *At any data release, let $\mathcal{AG}_i(o)$ be the anonymized group that contains the record for individual o in published table T_i^* .*

For the sake of clarity, if the context is clear, we omit the subscript and denote $\mathcal{AG}_i(o)$ by $\mathcal{AG}(o)$.

DEFINITION 4 (POSSIBLE WORLD FOR o). *Given a possible world $TS = \{T_1^p, T_2^p, \dots, T_k^p\}$ for $\{T_1^*, T_2^*, \dots, T_k^*\}$. Let us extract the tuples in each T_i^p that correspond to the tuples in the anonymized group $\mathcal{AG}_i(o)$ (containing individual o in T_i^*) to form table $T_i^p(o)$. Then, the series of smaller tables, denoted by $TS(o)$ which is equal to $\{T_1^p(o), T_2^p(o), \dots, T_k^p(o)\}$, form a possible world for $\mathcal{AG}_1(o), \dots, \mathcal{AG}_k(o)$. We also say that that $TS(o)$ is a possible world for o for $\{T_1^*, T_2^*, \dots, T_k^*\}$.*

For example, Figure 3 shows all the possible worlds for G_1 and G_2 for o_2 in the published tables shown in Figure 2(a) and Figure 2(b). Note that in the above definition, if o does not appear in a table T_i , then $T_i^p(o)$ is an empty table.

4.1 Possible World Analysis

Since the sensitive values are transient and we do not assume any additional knowledge about the data linkage, the assignment of sensitive values to individuals in groups other than $\mathcal{AG}(o)$ are independent of the assignment to the individuals in $\mathcal{AG}(o)$. Hence, we arrive at the following lemma.

LEMMA 2. *The value of $p(o, s, k)$ can be derived based on the analysis of the possible worlds for o .*

The above lemma helps to greatly simplify the analysis of the privacy breach by considering only $\mathcal{AG}(o)$ in each data release. In the following, we may refer to a possible world for o simply as a possible world.

Consider an anonymized group $\mathcal{AG}(o)$ in T_j for individual o . Let n_j be the size $|\mathcal{AG}(o)|$. Let $n_{j,i}$ be the total number of tuples in $\mathcal{AG}(o)$ with sensitive value s_i for $i = 1, 2, \dots, |S|$. The total number of possible worlds for $\mathcal{AG}(o)$ can be derived by combinatorial analysis.

LEMMA 3 (NO. OF POSS. WORLDS FOR SINGLE TABLE). *The total number of possible worlds for the anonymized group $\mathcal{AG}(o)$ in a single published table T_j^* is equal to*

$$W_j = \frac{n_j!}{\prod_{i=1}^{|S|} n_{j,i}!}$$

For example, consider an anonymized group of size 4 containing two s_1 values, one s_2 value and one s_3 value in T_j^* . Then, W_j is equal to $\frac{4!}{2! \times 1! \times 1!} = 12$.

4.2 Breach Probability

Recall that our objective is to compute $p(o, s, k)$ which involves two major components, namely $W_{link}(o, s, k)$ and $W_{total,k}$. In the following, we will describe how we obtain the values of these two components.

By Lemma 3, the total number of possible worlds for o in the published tables $T_1^*, T_2^*, \dots, T_k^*$, denoted by $W_{total,k}$, is equal to

$$W_{total,k} = \prod_{j=1}^k W_j = \prod_{j=1}^k \frac{n_j!}{\prod_{i=1}^{|S|} n_{j,i}!} \quad (2)$$

Next, we will describe how to obtain the formula for $W_{link}(o, s, k)$. Without loss of generality, we consider the privacy protection for an arbitrary sensitive value $s = s_1$. The following analysis applies for each sensitive value.

Note that, for any arbitrary sensitive value s_1 , we have the following.

$$W_{total,k} = W_{link}(o, s_1, k) + \overline{W}_{link}(o, s_1, k)$$

where $\overline{W}_{link}(o, s_1, k)$ is the total number of possible worlds where o is not linked to s_1 in all k published tables, namely $T_1^p, T_2^p, \dots, T_k^p$. Thus,

$$W_{link}(o, s_1, k) = W_{total,k} - \overline{W}_{link}(o, s_1, k)$$

Next, we will show how we derive $\overline{W}_{link}(o, s_1, k)$. Let $\theta(o, s_1, j)$ be the total number of possible worlds for table T_j^p (treated as a singleton table series) that o is not linked to s_1 .

Consider a possible table T_j^p where o is not linked to s_1 . Since o is not linked to s_1 in T_j^p , o is linked to a sensitive value s_q where $q \neq 1$ in T_j^p . The number of possible worlds for T_j^p where o is linked to s_q in T_j^p is equal to

$$W_{sq,j} = \frac{(n_j - 1)!}{(n_{j,q} - 1)! \prod_{i=1, i \neq q}^{|S|} n_{j,i}!}$$

By considering all sensitive values s_q where $q \in [2, |S|]$, the total number of possible worlds for T_j^p where o is not linked to s_1 (i.e., $\theta(o, s_1, j)$) is equal to

$$\begin{aligned} \sum_{q=2}^{|S|} W_{sq,j} &= \sum_{q=2}^{|S|} \frac{(n_j - 1)!}{(n_{j,q} - 1)! \prod_{i=1, i \neq q}^{|S|} n_{j,i}!} \\ &= \sum_{q=2}^{|S|} \frac{(n_j - 1)! n_{j,q}}{\prod_{i=1}^{|S|} n_{j,i}!} \\ &= \frac{(n_j - 1)!}{\prod_{i=1}^{|S|} n_{j,i}!} \sum_{q=2}^{|S|} n_{j,q} \end{aligned}$$

Consider $\overline{W}_{link}(o, s_1, k)$

$$\begin{aligned} &= \prod_{j=1}^k \theta(o, s_1, j) \\ &= \prod_{j=1}^k \left[\frac{(n_j - 1)!}{\prod_{i=1}^{|S|} n_{j,i}!} \sum_{q=2}^{|S|} n_{j,q} \right] \\ &= \left(\prod_{j=1}^k \frac{(n_j - 1)!}{\prod_{i=1}^{|S|} n_{j,i}!} \right) \left(\prod_{j=1}^k \sum_{q=2}^{|S|} n_{j,q} \right) \\ &= \left(\prod_{j=1}^k \frac{(n_j - 1)!}{\prod_{i=1}^{|S|} n_{j,i}!} \right) \left(\prod_{j=1}^k (n_j - n_{j,1}) \right) \end{aligned}$$

From Equation (1),

$$\begin{aligned} &p(o, s_1, k) \\ &= \frac{W_{link}(o, s_1, k)}{W_{total,k}} \\ &= \frac{W_{total,k} - \overline{W}_{link}(o, s_1, k)}{W_{total,k}} \\ &= \frac{\prod_{j=1}^k \frac{n_j!}{\prod_{i=1}^{|S|} n_{j,i}!} - \left(\prod_{j=1}^k \frac{(n_j - 1)!}{\prod_{i=1}^{|S|} n_{j,i}!} \right) \left(\prod_{j=1}^k (n_j - n_{j,1}) \right)}{\prod_{j=1}^k \frac{n_j!}{\prod_{i=1}^{|S|} n_{j,i}!}} \\ &= \frac{\prod_{j=1}^k n_j - \prod_{j=1}^k (n_j - n_{j,1})}{\prod_{j=1}^k n_j} \end{aligned}$$

LEMMA 4 (CLOSED FORM OF $p(o, s_1, k)$).

$$p(o, s_1, k) = \frac{\prod_{j=1}^k n_j - \prod_{j=1}^k (n_j - n_{j,1})}{\prod_{j=1}^k n_j} \quad (3)$$

From Equation (1), $p(o, s_1, k)$ is defined with a conceptual terms with the total number of possible worlds. Lemma 4 gives a closed form of $p(o, s_1, k)$. Given the information of n_j (i.e., the size of the anonymized group in the j -th table) and $n_{j,1}$ (i.e., the number of tuples in the anonymized group with sensitive value s_1 in the j -th table), we can calculate $p(o, s_1, k)$ with its closed form directly.

EXAMPLE 1 (TWO-TABLE ILLUSTRATION). Consider that we want to protect the linkage between an individual and a sensitive value s_1 . Suppose o appears in both published tables T_1^* and T_2^* . Let $\mathcal{AG}_1(o)$ and $\mathcal{AG}_2(o)$ be the anonymized groups in T_1^* and T_2^* containing o . Suppose both $\mathcal{AG}_1(o)$ and $\mathcal{AG}_2(o)$ are linked to s_1 .

By the notation adopted in this paper, n_k is the size of $\mathcal{AG}_k(o)$ and $n_{k,1}$ is the total number of tuples in $\mathcal{AG}_k(o)$ with sensitive value s_1 .

By Lemma 4, we have

$$\begin{aligned} p(o, s_1, k) &= \frac{n_1 n_2 - (n_1 - n_{1,1})(n_2 - n_{2,1})}{n_1 n_2} \\ &= \frac{n_{2,1} n_1 + n_{1,1} n_2 - n_{1,1} n_{2,1}}{n_1 n_2} \end{aligned}$$

□

EXAMPLE 2 (RUNNING EXAMPLE). In our running example as shown in Figure 2, consider the second individual o_2 and a sensitive value “chlamydia”. We know that $n_1 = n_2 = 2$. Suppose s_1 is “chlamydia”. Thus, $n_{1,1} = n_{2,1} = 1$. With respect to the published tables as shown in Figure 2, according to the formula derived in Example 1,

$$p(o_2, s_1, 2) = \frac{1 \times 2 + 1 \times 2 - 1 \times 1}{2 \times 2} = \frac{3}{4}$$

which is greater than $1/2$ (the desired threshold).

However, if we publish tables as shown in Figure 4, then $n_1 = n_2 = 4$ and $n_{1,1} = n_{2,1} = 1$.

$$p(o_2, s_1, 2) = \frac{1 \times 4 + 1 \times 4 - 1 \times 1}{4 \times 4} = \frac{7}{16}$$

which is smaller than $1/2$.

In this example, we observe that, since the published tables as shown in Figure 4 have a larger anonymized group size (compared with the published tables as shown in Figure 2), $p(o, s_1, 2)$ is smaller. \square

In this paper, we aim to publish table T_k^* like Figure 4 at each time point k such that $p(o, s, k) \leq 1/\ell$ for each individual o and each sensitive value s .

From Example 2, we observe that a larger anonymized group size reduces the breach probability that individual o is linked to sensitive value s_1 in the past. However, the anonymized group size alone cannot reduce the breach probability. Consider that an anonymized group in published table T_k^* contains all sensitive values s_1 , instead of distinct sensitive values. Even though this anonymized group is larger, if it still contains all sensitive values s_1 , it is easy to verify that an individual o in this anonymized group must be linked to s_1 in this table T_k^* .

In fact, the breach probability is determined by the *anonymized group size ratio*. The anonymized group size ratio is equal to the anonymized group size divided by the total number of tuples in this anonymized group with sensitive value s_1 . In Example 2, since all sensitive values are distinct in an anonymized group (i.e., the total number of tuples in this anonymized group with sensitive value s_1 is equal to 1), the anonymized group size ratio is equal to the anonymized group size. In the next section, we will show that the larger anonymized group size ratio can reduce the probability.

5. THEORETICAL PROPERTIES

In the previous section, we describe that a larger anonymized group ratio can reduce the breach probability. In this section, we will first study some properties of our problem, including a *minimum* anonymized group ratio for global privacy guarantee, and then a monotonicity property that can be useful in data anonymization.

5.1 Minimum \mathcal{AG} size Ratio

Recall that n_k is the anonymized group (\mathcal{AG}) size and $n_{k,1}$ is the number of tuples in the anonymized group with sensitive value s_1 . In the following, we will derive the minimum anonymized group size ratio $\frac{n_k}{n_{k,1}}$ for privacy protection under the global guarantee.

THEOREM 1. *Let k be an integer greater than 1. Suppose the anonymized group in T_k^* containing individual o is linked to s_1 . $p(o, s_1, k) \leq 1/\ell$ if and only if*

$$\frac{n_k}{n_{k,1}} \geq \frac{\ell \prod_{j=1}^{k-1} (n_j - n_{j,1})}{\ell \prod_{j=1}^{k-1} (n_j - n_{j,1}) - (\ell - 1) \prod_{j=1}^{k-1} n_j} \quad (4)$$

Proof: By Lemma 4, $p(o, s, k)$ is equal to $\frac{\prod_{j=1}^k n_j - \prod_{j=1}^k (n_j - n_{j,1})}{\prod_{j=1}^k n_j}$.

$$p(o, s_1, k) \leq 1/\ell$$

$$\begin{aligned} & \Leftrightarrow \frac{\prod_{j=1}^k n_j - \prod_{j=1}^k (n_j - n_{j,1})}{\prod_{j=1}^k n_j} \leq \frac{1}{\ell} \\ & \Leftrightarrow \frac{n_k \prod_{j=1}^{k-1} n_j - (n_k - n_{k,1}) \prod_{j=1}^{k-1} (n_j - n_{j,1})}{n_k \prod_{j=1}^{k-1} n_j} \leq \frac{1}{\ell} \\ & \Leftrightarrow \frac{\prod_{j=1}^{k-1} n_j - (1 - \frac{n_{k,1}}{n_k}) \prod_{j=1}^{k-1} (n_j - n_{j,1})}{\prod_{j=1}^{k-1} n_j} \leq \frac{1}{\ell} \\ & \Leftrightarrow \frac{n_{k,1}}{n_k} \leq 1 - \frac{\ell \prod_{j=1}^{k-1} n_j - \prod_{j=1}^{k-1} n_j}{\ell \prod_{j=1}^{k-1} (n_j - n_{j,1})} \\ & \Leftrightarrow \frac{n_k}{n_{k,1}} \geq \frac{\ell \prod_{j=1}^{k-1} (n_j - n_{j,1})}{\ell \prod_{j=1}^{k-1} (n_j - n_{j,1}) - (\ell - 1) \prod_{j=1}^{k-1} n_j} \end{aligned}$$

\square

From the above, for any $k > 1$, we can see that the value of $\frac{n_k}{n_{k,1}}$ should be lower bounded by the value of

$$\underline{n}(k) = \frac{\ell \prod_{j=1}^{k-1} (n_j - n_{j,1})}{\ell \prod_{j=1}^{k-1} (n_j - n_{j,1}) - (\ell - 1) \prod_{j=1}^{k-1} n_j}$$

We define $\underline{n}(k) = \ell$ when $k = 1$.

EXAMPLE 3 (RUNNING EXAMPLE). From Example 2, we know that the published tables shown in Figure 4 satisfy the privacy requirement (i.e., $p(o, s, k) \leq 1/\ell$ where $k = 2$ and $\ell = 2$). At time $k = 3$, we want to publish a new table T_3^* from a raw table T_3 which contain o_2 .

Suppose we will put o_2 in the anonymized group $\mathcal{AG}_3(o_2)$ in T_3^* which is linked to s_1 where s_1 is chlamydia. By Theorem 1, when $k = 3$, the R.H.S. of Equation (4) becomes

$$\begin{aligned} & \frac{\ell(n_1 - n_{1,1})(n_2 - n_{2,1})}{\ell(n_1 - n_{1,1})(n_2 - n_{2,1}) - (\ell - 1)n_1 n_2} \\ & = \frac{2(4 - 1)(4 - 1)}{2(4 - 1)(4 - 1) - (2 - 1) \cdot 4 \cdot 4} \\ & = 9 \end{aligned}$$

which is the minimum anonymized group size ratio $\frac{n_3}{n_{3,1}}$ in the published table T_3^* . Suppose $\mathcal{AG}_3(o)$ contains only one occurrence of s_1 . Then, the size of the anonymized group $\mathcal{AG}_3(o)$ should be at least 9 so that $p(o, s_1, 3) \leq 1/2$. \square

We have the following corollary when the inequality in Theorem 1 becomes an equality.

COROLLARY 2. $\frac{n_k}{n_{k,1}} = \underline{n}(k)$ if and only if $p(o, s_1, k) = 1/\ell$.

When a record for individual o appears in a data release T_i and in the published data T_i^* , the anonymized group containing o has no relation to sensitive value s , then intuitively, this release should not have any impact on the privacy protection of o linking to s . This is formally stated in the following lemma.

LEMMA 5. *If the anonymized group in T_k^* containing o is not linked to s_1 , then $p(o, s_1, k) = p(o, s_1, k - 1)$.*

Proof: Since the anonymized group in T_k^* containing o is not linked to s_1 , we know that o is linked to s_1 in one of the first $(k-1)$ -th published tables. Thus,

$$W_{link}(o, s_1, k) = W_k \times W_{link}(o, s_1, k-1)$$

Thus, we have

$$\begin{aligned} p(o, s, k) &= \frac{W_{link}(o, s_1, k)}{W_{total, k}} \\ &= \frac{W_k \times W_{link}(o, s_1, k-1)}{\prod_{j=1}^k W_j} \quad (\text{From Equation (2)}) \\ &= \frac{W_{link}(o, s_1, k-1)}{W_{total, k-1}} \\ &= p(o, s, k-1) \end{aligned}$$

□

Thus, $\frac{n_k}{n_{k,1}}$ can be equal to any real number and does not affect the value of $p(o, s, k)$ in this case.

Suppose a published table T_k^* contains o and we need to generate an anonymized group G containing o . Note that the size of the anonymized group G is n_k and the number of tuples in G with sensitive value s_i is equal to $n_{k,i}$ for $i = [1, |S|]$. Without loss of generality, suppose we want to protect the privacy linkage between an individual o and a sensitive value s_1 . From Theorem 1 and Lemma 5, we can determine the minimum value of $\frac{n_k}{n_{k,1}}$ for generating an anonymized group G . From Theorem 1, if G contains s_1 , in order to guarantee $p(o, s_1, k) \leq 1/\ell$, we have to set the value of n_k to satisfy

$$\frac{n_k}{n_{k,1}} \geq \tilde{n}(k)$$

From Lemma 5, if G does not contain s_1 , any value of $\frac{n_k}{n_{k,1}}$ will not affect the privacy related to o and s_1 .

Although Theorem 1 suggests that if we set the value of $\frac{n_k}{n_{k,1}}$ at least $\tilde{n}(k)$, then $p(o, s_1, k) \leq 1/\ell$. However, suppose we set this value *exactly* equal to $\tilde{n}(k)$, although we can guarantee $p(o, s_1, k) \leq 1/\ell$ for the k published tables, there will be a privacy breach (i.e., $p(o, s_1, k') > 1/\ell$) for any additional *future* published tables in which an anonymized group containing o is linked to s_1 . This is a result of the following lemma.

THEOREM 2. *Consider that we published $k-1$ tables where an anonymized group in T_{k-1}^* containing o is linked to s_1 . Suppose we are to publish T_k^* where an anonymized group in T_k^* containing o is also linked to s_1 . If $\frac{n_{k-1}}{n_{k-1,1}} = \tilde{n}(k-1)$, then $p(o, s_1, k) > 1/\ell$.*

5.2 Monotonicity

Monotonicity is a useful property for some anonymization process where the resulting anonymization groups are constructed in a bottom-up manner, merging smaller groups that violates the privacy requirement into bigger groups which may guarantee privacy. It is also useful when the anonymization is top-down, splitting bigger groups into smaller ones as long as the privacy guarantee holds.

Consider the privacy protection for the linkage of an individual o to a sensitive value s_1 . From Lemma 5, we know that $p(o, s_1, k)$ is *independent* of data releases in which any anonymized group containing o (in the published tables) are

not linked to s_1 . Hence, in the following, we consider the worst-case scenario where in all releases whenever there exists an anonymized group containing o (in a published table), o is linked to s_1 .

The monotonicity property is described as follows.

THEOREM 3 (MONOTONICITY). *$p(o, s_1, k)$ is strictly decreasing when $\frac{n_k}{n_{k,1}}$ increases.*

The proof is given in the appendix. Note that $n_k/n_{k,1}$ is essentially the inverse of the proportion of s_1 tuples in the anonymized group. Therefore, when a bigger group that satisfies the privacy requirement is split into smaller ones, if the proportion of s_1 tuples in the small group containing o is not increased, then $p(o, s_1, k)$ is not increased. Conversely if a small group violates the privacy guarantee, merging it with another group may decrease the proportion of s_1 tuples and thus $p(o, s, k)$ may be decreased.

An anonymized group \mathcal{AG} is said to violate the global guarantee if there exists an individual $o \in \mathcal{AG}.I$ and a sensitive value $s \in \mathcal{AG}.S$ such that $p(o, s, k) > 1/\ell$.

COROLLARY 3. *Consider an anonymized group \mathcal{AG} in the published table T_k^* which violates the global guarantee. If we partition \mathcal{AG} into a number of smaller groups, one of the smaller groups violates the global guarantee.*

Proof Sketch: Suppose $n_k/n_{k,1}$ is the size ratio for \mathcal{AG} . It is easy to see that one of the smaller groups has the size ratio smaller than $n_k/n_{k,1}$. By Theorem 3, $p(o, s, k)$ increases. Since \mathcal{AG} violates the global guarantee (i.e., $p(o, s, k) > 1/\ell$), the smaller group also violates the global guarantee (i.e., $p(o, s, k) > 1/\ell$). □

6. ANONYMIZATION

In previous sections, we have observed that, by choosing a proper size of an anonymized group, the global privacy guarantee can be achieved. In general, a size above a certain threshold size can be chosen. However, setting a size equal to the threshold size will make future anonymization infeasible (see Theorem 2). Therefore, it is necessary to choose a size that is greater than the threshold. The increase in size however, would lead to a decrease in the utility of the data. Hence, a question will be how to pick a smallest size that can maintain the global guarantee.

In this section, we show that if we are given a bound on the number of releases where an individual o may be linked to a sensitive value s , then we can devise a strategy to minimize the maximum anonymization group size. We also propose another strategy which aims to reduce the anonymized group size on average.

6.1 Constant-Ratio Strategy

In database related problems, one can typically derive effective mechanisms based on the characteristics of the data itself. In our problem scenario, a data publisher has at his/her disposal the statistical information of the data collections. For example, consider the medical database. The statistics can point to the expected frequency of an individual contracting a certain disease over his or her lifespan. With such information, one can set an estimated bound on the number of data releases that a person may indeed be linked to

ℓ	2	2	2	5	10	2	5	10
k'	2	5	20	20	20	10	10	10
\tilde{n}_c	3.44	7.75	29.41	90.13	190.33	15.10	45.35	95.42

Table 1: Values of \tilde{n}_c with selected values of ℓ and k'

the disease. With this knowledge, one can adopt a constant-ratio strategy which we shall show readily can minimize the maximum size of the corresponding anonymized groups.

Constant-ratio strategy makes sure that the size of anonymized groups $\mathcal{AG}(o)$ for individual o containing s_1 divided by the number of occurrences of s_1 remain unchanged over a number of data releases. Formally, given an integer k' for the number of data releases, for $i \in [1, k']$,

$$\frac{n_{o_i}}{n_{o_i,1}} = \tilde{n}_c$$

where \tilde{n}_c is a positive real number constant, and o_i is a timestamp for the i -th release where both o and s_1 appear. For the sake of simplicity, we set $\tilde{n}_c = \frac{n}{n_s}$ where n and n_s are positive integer constants where $n_s \leq n$.

k' corresponds to the total number of possible releases in the future. In other words, during data publishing, the data publisher expects to publish k' table for this data. With this given parameter k' , we can calculate n and n_s such that $\frac{n_{o_i}}{n_{o_i,1}}$ remain unchanged when i changes.

In order to make sure that $p(o, s_1, j) \leq 1/\ell$ for any $j \in [1, k']$, we need to protect $p(o, s_1, k') \leq 1/\ell$. In the following, we consider $p(o, s_1, k')$ which is equal to

$$\begin{aligned} \frac{\prod_{j=1}^{k'} n_j - \prod_{j=1}^{k'} (n_j - n_{j,1})}{\prod_{j=1}^{k'} n_j} &\leq \frac{1}{\ell} \\ n^{k'} - (n - n_s)^{k'} &\leq n^{k'} \times \frac{1}{\ell} \\ 1 - (1 - \frac{n_s}{n})^{k'} &\leq \frac{1}{\ell} \\ \frac{n_s}{n} &\leq 1 - (1 - \frac{1}{\ell})^{1/k'} \\ \frac{n}{n_s} &\geq 1/[1 - (1 - \frac{1}{\ell})^{1/k'}] \end{aligned}$$

Let $\tilde{n}_c = 1/[1 - (1 - \frac{1}{\ell})^{1/k'}]$.

Table 1 shows the values of \tilde{n}_c with selected values of ℓ and k' . When ℓ increases, \tilde{n}_c increases. When k' increases, \tilde{n}_c also increases.

It remains to show that the constant-ratio strategy indeed can lead to data publishing that minimizes the maximum anonymized group sizes. First, we define this property more formally.

DEFINITION 5 (MIN-MAX OPTIMIZATION). *An anonymization for serial data publishing is min-max optimal if the maximum anonymized group size among the anonymized groups containing individual o and sensitive value s_1 for any given o and s_1 over all data releases is minimized.*

THEOREM 4 (OPTIMALITY). *The constant-ratio strategy generates a min-max optimal solution for serial data publishing.*

Proof: Let N be the set of anonymized group sizes in the k' published tables where these anonymized groups contain

o and are linked to s_1 . That is, $N = \{n_1, n_2, \dots, n_{k'}\}$. Let $u(N) = \max_{n_i \in N} n_i$. Let N_a be the set of anonymized group sizes in the k' published tables generated by strategy a .

Let $p(o, s_1, k'|a)$ be $p(o, s_1, k')$ with respect to strategy a . Let A be the set of all possible strategies a such that, with the published tables with strategy a , $p(o, s_1, k'|a) \leq 1/\ell$. Suppose a_o is the constant-ratio strategy. We will prove that this strategy can obtain an optimal value of $u(N)$. That is,

$$u(N_{a_o}) = \min_{a \in A} \{u(N_a)\}$$

We prove by contradiction. Consider that the strategy a_o generates $N_{a_o} = \{n_1, n_2, \dots, n_{k'}\}$. By Corollary 2, it is easy to verify that $p(o, s_1, k'|a_o) = 1/\ell$.

Suppose there exists a strategy $a' \neq a_o$ which generates $N_{a'} = \{n'_1, n'_2, \dots, n'_{k'}\}$ such that $u(N_{a'}) < u(N_{a_o})$ and $p(o, s_1, k'|a') \leq 1/\ell$. We deduce that, for all $i \in [1, k']$,

$$n'_i < n_i$$

By Theorem 2, we know that, $p(o, s_1, k'|a') > p(o, s_1, k'|a_o)$ (which is equal to $1/\ell$). We conclude that $p(o, s_1, k'|a') > 1/\ell$. Thus, privacy breach occurs, which leads to a contradiction. \square

Although the constant-ratio strategy generates a min-max optimal solution, the statistical information about the data should be known. For example, the constant-ratio strategy requires the priori knowledge about k' which is equal to the total number of possible releases in the future. If such information is unavailable, we can use the *geometric strategy* proposed in the next subsection where this strategy does not require the statistical information.

6.2 Geometric Strategy

Other than minimizing the maximum anonymized group size, another desirable utility criterion will be to minimize the average group size. In order to achieve this goal, we examine the probability of occurrences of anonymized groups for linking individuals o to a certain sensitive value s . From past data, there will be a distribution of the total number of releases where any given individual has contracted disease s . For example, if the maximum of such value is 10, some individuals may be linked to s 10 times in total, but most individuals may be linked to s less than 10 times in total. Typically, the number of individuals that are linked to s for at least k releases will be greater than that for k'' releases where $k < k''$. Therefore, when choosing the sizes of the anonymized groups, it will reduce the average group size if we choose smaller sizes for the earlier releases where o is linked to s and bigger sizes for the later such releases. This is the essence of our next proposed strategy, namely, the *geometric strategy*.

With the geometric strategy, the anonymized group size will be equal to the minimum feasible value of $\tilde{n}(k)$ multiplied by a factor, α , at any time point k . This will be a growing value since the value of \tilde{n} will grow with k . Note that α must be greater than 1 since with $\alpha = 1$, the minimum feasible $\tilde{n}(k)$ will be used, and from Theorem 2, that will make future selection of group size infeasible. The value of α can be selected based on the estimated number of releases where an individual will be linked to s in total.

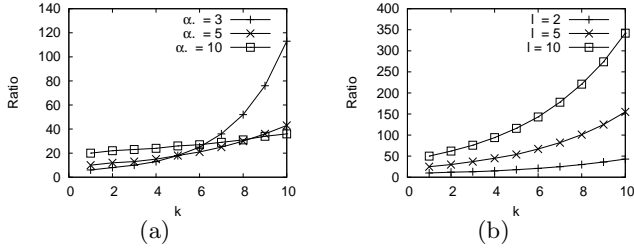


Figure 5: Effect of α and l on ratio $n_k/n_{k,1}$

Thus, with this strategy, with $j \geq 1$, we set

$$\frac{n_j}{n_{j,1}} = \alpha \cdot \tilde{n}(j)$$

Figure 5 shows how the values of $\frac{n_k}{n_{k,1}}$ increases with k . Figure 5(a) studies the effect of α (with l set to 2) and Figure 5(b) studies the effect of l (with α set to 5). When k increases, $\frac{n_k}{n_{k,1}}$ increases. When α is larger, although the initial value of $\frac{n_k}{n_{k,1}}$ is larger, the grow rate of $\frac{n_k}{n_{k,1}}$ is smaller. When l increases, $\frac{n_k}{n_{k,1}}$ increases. Figure 5(a) shows that $\alpha = 5$ and 10 are better choices than $\alpha = 3$ since the increase in the ratio is much slower. Note that these values are all pre-computable and it is easy to choose a suitable parameter by examining the pre-computed trends.

6.3 Discussion

In both of the above strategies, there may occur rare occasions where the required anonymized group size is not available in the given data set. As with previous works [30, 5], we handle the exceptional cases by data distortion. We can suppress the sensitive values of individuals when it is found that no feasible group size can maintain the global guarantee for privacy preservation. From our experimental results, such suppression has not been found needed.

Though our discussion has been based on a single value for the sensitive attribute in each record, our results can be easily extended to the case where each record may contain a set of values for the sensitive attribute. The essential proportion of possible worlds where an individual is linked to a sensitive value would not be affected.

7. IMPLEMENTATION

In Section 6, we describe two strategies to determine the value of $\frac{n_k}{n_{k,1}}$ for privacy protection with respect to a sensitive value s_1 . In the following, we describe how to anonymize the table given the desired value of $\frac{n_k}{n_{k,1}}$.

Since the formula is based on the frequency that a tuple for individual o is linked to a sensitive value s in an anonymized group from published tables (by Theorem 1 and Lemma 5), we propose to keep a data structure, called *statistics file*, to store the sizes of the anonymized groups containing a record for individual o such that o is linked to a sensitive value s , denoted by $m(o, s)$. Consider an individual o and a sensitive value s . Let the anonymized groups containing o in T_1^*, T_2^* and T_3^* be G_1 (of size 3), G_2 (of size 5) and G_3 (of size 4), respectively. If G_1 and G_2 contain s but G_3 does not, $m(o, s)$ is equal to $\{3, 5\}$. Suppose there is another published table T_4^* which does not contain o . $m(o, s)$ is also equal to $\{3, 5\}$.

Given the statistics file, it is possible to adopt existing known anonymization methods to generate anonymized groups that satisfy the group size ratio requirement of interest. For example, we may use a bottom-up approach to grow the anonymized groups. Alternatively, we can use a top-down approach to keep breaking up large anonymized groups and stops when it begins to violate the group size ratio requirement.

8. EMPIRICAL STUDIES

All of our experiments have been performed on a Linux workstation with a 3.2Ghz CPU and 2 Giga-byte memory. Similar to [5], we deploy one public available real hospital database CDRMP¹. In the database, there are 8 tables: *Reports*, *Reactions*, *Drugs*, *ReportDrug*, *Ingredients*, *Outcome*, and *Druginvolve*. *Reports* consists of some patients' basic personal information. Therefore, we take it as the voter registration list. *Reactions* has a foreign key *PID* referring to the attribute *ID* in *Reports* and another attribute to indicate the person's disease. After removing tuples with missing values, *Reactions* has 105,420 tuples while *Reports* contains 40,478 different individuals. We take 10% least frequent sensitive values as transient sensitive values. There are totally 232 transient sensitive values.

Dynamic microdata table series $TS_{exp} = \{T_1, T_2, \dots, T_{20}\}$ is created from *Reactions*. We divide *Reactions* into 20 partitions of the same size, namely P_1, P_2, \dots, P_{20} . T_1 is set to P_1 . For each $i \in [2, 20]$, we generate T_i as follows. T_i is set to P_i initially. Then, we randomly select 20% of tuples in T_{i-1} and insert them into T_i . Then, in the resulting T_i , we randomly select 20% of tuples and change their values in the sensitive attribute according to the sensitive value distribution of all tuples in *Reaction* as follows. For each selected tuple t in the above step, we randomly pick a tuple t' in the original data *Reactions* and set the sensitive value of t in T_i to be the sensitive value of t' obtained in *Reaction*.

For our experiments, we have chosen a bottom-up anonymization algorithm [25] with a variation of involving the individuals that are present in the registration voter list but absent in the data release. Such individuals can be virtually included in an anonymized group and help to dilute the linkage probability of individuals to sensitive values in the group [18, 5]. This variation helps to improve the utility since a group can now consist of fewer records that are actually present in the data. A bottom-up approach is chosen because we find that typically the anonymized groups can be easily formed based on the smallest group sizes that satisfy the required group size ratios. In the constant-ratio strategy, the default value of k' is equal to 20. In the geometric strategy, the default value of α is set to 2.

We have tested our proposed method in terms of effectiveness and efficiency. For the evaluation of our method, we examine four different aspects: the average size of the anonymized groups in the published tables, the greatest size of the anonymized groups in the published tables, the utility of the published tables and the computation overheads.

For measuring the utility of the published data we compare query processing results on each anonymized table T_j^* and its corresponding microdata table T_j at each publishing round. We follow the literature conventions [28, 30, 26, 5] to measure

¹http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index_e.html

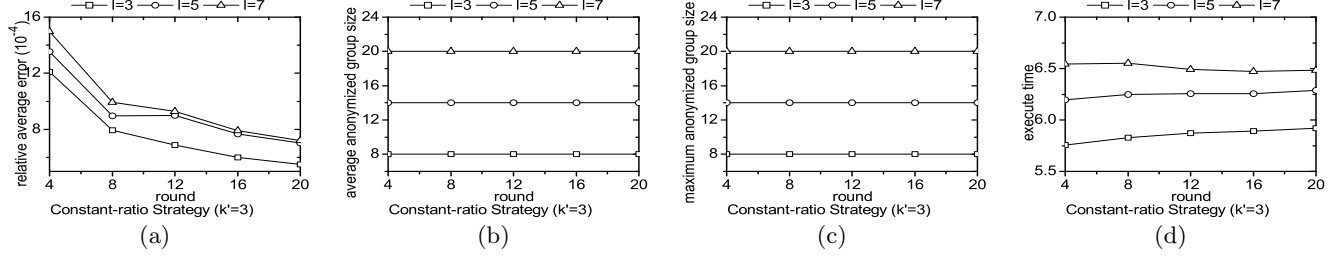


Figure 6: Effect of l (Constant-Ratio Strategy) where $k' = 3$

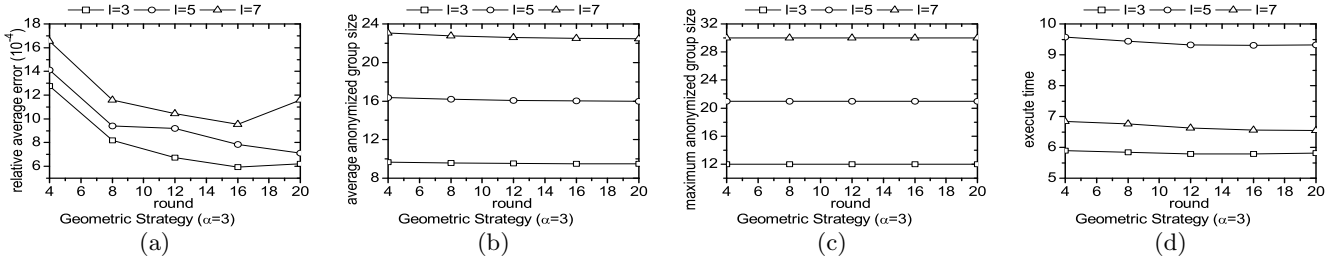


Figure 7: Effect of l (Geometric Strategy) where $\alpha = 3$

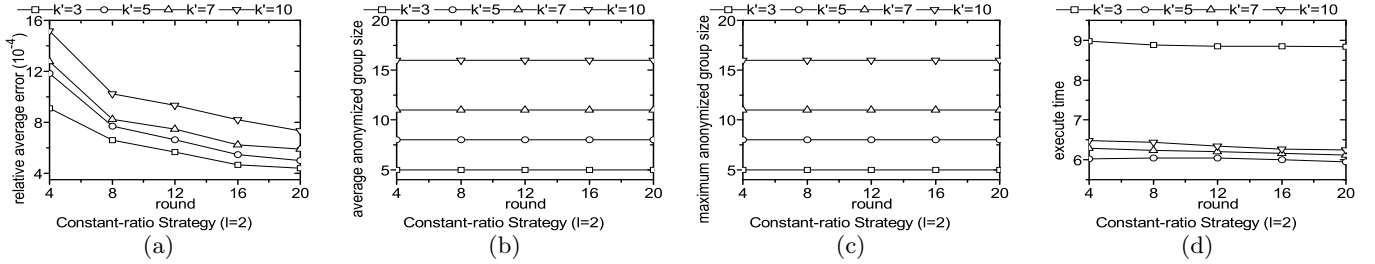


Figure 8: Effect of k' (Constant-Ratio Strategy) where $l = 2$

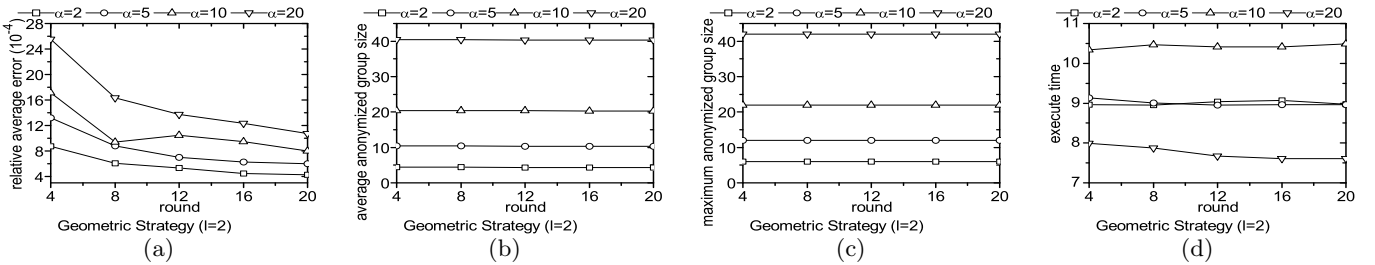


Figure 9: Effect of α (Geometric Strategy) where $l = 2$

the error by the relative error ratio in answering an aggregate query. All the published tables are evaluated one by one. For each evaluation, we perform 5,000 randomly generated range queries which follows the methodology in [30] on the microdata snapshot and its anonymized version, and then report the average relative error ratio.

We study the effect of variations in (1) the number of rounds, (2) the privacy requirement ℓ , (3) the parameter k' used in the constant-ratio strategy and (4) the parameter α used in the geometric strategy.

Effect of ℓ : Figure 6(a) shows that the average relative error of the constant-ratio strategy remains nearly unchanged when the number of rounds (or table releases) increases. As expected the error is larger with larger values of ℓ . In Figures 6(b) and (c), both the average anonymized group size and the maximum anonymized group size of the constant-ratio strategy keep nearly unchanged when we vary the number of rounds. Again as expected, the sizes increase with ℓ . Figure 6(d) shows that the execution time of the constant-ratio strategy keeps unchanged when there are more rounds. In the figure, when ℓ is larger, the execution time is larger. This is because we have to generate a larger anonymized group.

Figure 7 shows similar results for the geometric strategy with variation on the number of rounds. From Figures 7(a), (b) and (c) show that the error, the average anonymized group and the maximum anonymized group remains nearly unchanged when the number of rounds increases. In Figure 7(d), we cannot see a consistent trend when we vary ℓ . The execution time when $\ell = 3$ is the smallest. However, the execution time when $\ell = 7$ is smaller than that when $\ell = 5$. The execution time of the algorithm depends on two factors, namely the number of anonymized groups in the released tables and the sizes of the anonymized groups. Generating anonymized groups with larger sizes will increase the execution time. On the other hand, generating fewer anonymized groups will reduce the execution time. When $k' = 7$, since the factor of the total number of anonymized groups (i.e., fewer anonymized groups) outweighs the factor of the size of the anonymized group (i.e., larger anonymized group size), the execution time is smaller (compared with the case when $\ell = 5$).

Effect of k' : We study the input parameter of k' used in the constant-ratio strategy. In Figure 8, the average relative error, the average anonymized group size, the maximum group size and the execution time remains nearly unchanged when the number of rounds increases. The average relative error, the average anonymized group size and the maximum group size increases when k' increases as shown in Figures 8(a), (b) and (c). In Figure 8(d), we cannot observe a consistent trend of the execution time when k' increases. The reason is similar.

Effect of α : We also study the input parameter α for the geometric strategy. Similarly, Figure 9 shows that the number of rounds does not have a significant impact on the average relative error, the average anonymized group size, the maximum anonymized size and the execution time. Figures 9(a), (b) and (c) show that, when α increases, the average relative error, the average anonymized group size and the maximum anonymized size increases. There is no consistent trend for the execution time when we vary α as shown in Figures 9(d).

The reasons are similar to that in the study with the effect of ℓ .

Overall, our proposed methods are very efficient and introduce very small querying error. It shows that our method can provide the global guarantee on individual privacy as well as maintain high utility in the published data.

9. CONCLUSION

In this paper, we propose a new criterion of *global guarantee* for privacy preserving data publishing. This guarantee corresponds to a basic requirement of individual privacy where the probability of linking an individual to a sensitive value in one or more data releases is bounded. We show that global guarantee is a stronger privacy requirement than localized guarantee which has been adopted in previous works. We derive some theoretical results on this problem and discover that the anonymized group size is an important factor in privacy protection. According to the anonymized group sizes, we propose two strategies for anonymization. Our empirical study shows that these techniques are highly feasible and generate data publication of high utility.

There are some promising future directions. In this paper, we study the global guarantee for transient sensitive values, meaning that the values can change freely. As a future plan, we will study the global guarantee when both transient sensitive values and permanent sensitive values are present. Permanent sensitive values are studied in [5] and refer to values that will be permanently linked to an individual once it is linked to that individual. Intuitively, we can combine the technique here and that in [5] by forming the HD-compositions for holders and decoys, as well as forming anonymized groups based on the proper group size determined by our strategies here for taking care of the transient values. However, the details are left for future studies. Another direction is to extend the problem with the consideration of other background knowledge.

10. REFERENCES

- [1] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. In *EDBT*, 2004.
- [2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, 2005.
- [3] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, 2005.
- [4] E. Bertino, B.C. Ooi, Y. Yang, and R. Deng. Privacy and ownership preserving of outsourced medical data. In *ICDE*, 2005.
- [5] Y. Bu, A. W.-C. Fu, R. C.-W. Wong, L. Chen, and J. Li. Privacy preserving serial data publishing by role composition. In *VLDB*, 2008.
- [6] J. Byun, Y. Sohn, E. Bertino, and N. Li. Secure anonymization for incremental datasets. In *Secure Data Management*, pages 48–63, 2006.
- [7] Y. Du, T. Xia, Y. Tao, D. Zhang, and F. Zhu. On multidimensional k-anonymity with local recoding generalization. In *ICDE*, 2007.
- [8] B. C. M. Fung, K. Wang, A. Fu, and J. Pei. Anonymity for continuous data publishing. In *EDBT*, 2008.
- [9] G. Ghinita, Y. Tao, and P. Kalnis. On the anonymization of sparse high-dimensional data. In *ICDE*, 2008.
- [10] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, 2002.
- [11] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD*, 2006.

- [12] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, 2006.
- [13] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD*, 2005.
- [14] N. Li and T. Li. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In *ICDE*, 2007.
- [15] T. Li and N. Li. Injector: Mining background knowledge for data anonymization. In *ICDE*, 2008.
- [16] A. Machanavajjhala, J. Gehrke, and D. Kifer. *l*-diversity: privacy beyond *k*-anonymity. In *ICDE*, 2006.
- [17] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, 2004.
- [18] M. Nergiz, M. Atzori, and C.W. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD*, 2007.
- [19] J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang. Maintaining k-anonymity against incremental updates. In *SSDBM*, 2007.
- [20] V. Rastogi, D. Suci, and S. Hong. The boundary between privacy and utility in data publishing. In *VLDB*, 2007.
- [21] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowledge based systems*, 10(5), 2002.
- [22] Y. Tao, X. Xiao, J. Li, and D. Zhang. On anti-corruption privacy preserving publication. In *ICDE*, 2008.
- [23] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *EDBT*, 2008.
- [24] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In *ICDM05*, 2005.
- [25] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, 2004.
- [26] R.C.W. Wong, A. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, 2007.
- [27] R.C.W. Wong, J. Li, A. Fu, and K. Wang. (alpha, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In *KDD*, 2006.
- [28] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, 2006.
- [29] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD*, 2006.
- [30] X. Xiao and Y. Tao. *m*-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, 2007.
- [31] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu. Utility-based anonymization using local recoding. In *KDD*, 2006.
- [32] Y. Xu, K. Wang, A. W.-C. Fu, and P.S. Yu. Anonymizing transaction databases for publication. In *KDD*, 2008.
- [33] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on aononymized tables. In *ICDE*, 2007.

11. APPENDIX

Here we give the proofs of some of the lemmas and theorems listed in the previous sections.

Theorem 2: *Consider that we published $k - 1$ tables where an equivalence class in T_{k-1}^* containing t is linked to s_1 . Suppose we are to publish T_k^* where an equivalence class in T_k^* containing o is also linked to s_1 . If $\frac{n_{k-1}}{n_{k-1,1}} = \tilde{n}(k - 1)$, then $p(o, s_1, k) > 1/\ell$.*

Proof:

$$\begin{aligned} \frac{n_{k-1}}{n_{k-1,1}} &= \tilde{n}(k - 1) \\ &= \frac{\ell \prod_{j=1}^{k-2} (n_j - n_{j,1})}{\ell \prod_{j=1}^{k-2} (n_j - n_{j,1}) - (\ell - 1) \prod_{j=1}^{k-2} n_j} \\ &= \frac{\ell \prod_{j=1}^{k-1} (1 - \frac{n_{j,1}}{n_j})}{\ell \prod_{j=1}^{k-2} (1 - \frac{n_{j,1}}{n_j}) - (\ell - 1)} \end{aligned}$$

Let $P = \prod_{j=1}^{k-1} (1 - \frac{n_{j,1}}{n_j})$. We have

$$\frac{n_{k-1}}{n_{k-1,1}} = \frac{\ell P}{\ell P - \ell + 1}$$

Consider

$$\begin{aligned} p(o, s_1, k) &= \frac{\prod_{j=1}^k n_j - \prod_{j=1}^k (n_j - n_{j,1})}{\prod_{j=1}^k n_j} \\ &= 1 - \prod_{j=1}^k (1 - \frac{n_{j,1}}{n_j}) \\ &= 1 - (1 - \frac{n_{k,1}}{n_k}) (1 - \frac{n_{k-1,1}}{n_{k-1}}) \prod_{j=1}^{k-2} (1 - \frac{n_{j,1}}{n_j}) \\ &= 1 - (1 - \frac{n_{k,1}}{n_k}) (1 - \frac{\ell P}{\ell P - \ell + 1}) P \\ &= 1 - (1 - \frac{n_{k,1}}{n_k}) (1 - \frac{1}{\ell}) \\ &= \frac{1}{\ell} + \frac{n_{k,1}}{n_k} (1 - \frac{1}{\ell}) \\ &> \frac{1}{\ell} \end{aligned}$$

That is, $p(o, s_1, k) > 1/\ell$. □

Theorem 3 (MONOTONICITY). *$p(o, s_1, k)$ is strictly decreasing when $\frac{n_k}{n_{k,1}}$ increases.*

Proof:

$$\begin{aligned} p(o, s_1, k) &= \frac{\prod_{j=1}^k n_j - \prod_{j=1}^k (n_j - n_{j,1})}{\prod_{j=1}^k n_j} \\ &= 1 - \prod_{j=1}^k (1 - \frac{n_{j,1}}{n_j}) \\ &= 1 - (1 - \frac{n_{k,1}}{n_k}) \prod_{j=1}^{k-1} (1 - \frac{n_{j,1}}{n_j}) \end{aligned}$$

If $\frac{n_k}{n_{k,1}}$ increases, the above equation decreases. □